Biomedical Signal Processing and Control 14 (2014) 197–205

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Analyzing respiratory effort amplitude for automated sleep stage classification

Xi Long[a,b,*], Jérôme Foussier[c], Pedro Fonseca[a,b], Reinder Haakma[b], Ronald M. Aarts[a,b]

[a] Department of Electrical Engineering, Eindhoven University of Technology, The Netherlands
[b] Personal Health Group, Philips Group Innovation Research, The Netherlands
[c] Philips Chair for Medical Information Technology (MedIT), RWTH Aachen University, Germany

ARTICLE INFO

Article history:
Received 6 May 2014
Received in revised form 10 July 2014
Accepted 1 August 2014
Available online 27 August 2014

Keywords:
Respiratory effort amplitude
Signal calibration
Feature extraction
Sleep stage classification

ABSTRACT

Respiratory effort has been widely used for objective analysis of human sleep during bedtime. Several features extracted from respiratory effort signal have succeeded in automated sleep stage classification throughout the night such as variability of respiratory frequency, spectral powers in different frequency bands, respiratory regularity and self-similarity. In regard to the respiratory amplitude, it has been found that the respiratory depth is more irregular and the tidal volume is smaller during rapid-eye-movement (REM) sleep than during non-REM (NREM) sleep. However, these physiological properties have not been explicitly elaborated for sleep stage classification. By analyzing the respiratory effort amplitude, we propose a set of 12 novel features that should reflect respiratory depth and volume, respectively. They are expected to help classify sleep stages. Experiments were conducted with a data set of 48 sleepers using a linear discriminant (LD) classifier and classification performance was evaluated by overall accuracy and Cohen's Kappa coefficient of agreement. Cross validations (10-fold) show that adding the new features into the existing feature set achieved significantly improved results in classifying wake, REM sleep, light sleep and deep sleep (Kappa of 0.38 and accuracy of 63.8%) and in classifying wake, REM sleep and NREM sleep (Kappa of 0.45 and accuracy of 76.2%). In particular, the incorporation of these new features can help improve deep sleep detection to more extent (with a Kappa coefficient increasing from 0.33 to 0.43). We also revealed that calibrating the respiratory effort signals by means of body movements and performing subject-specific feature normalization can ultimately yield enhanced classification performance.

1. Introduction

According to the rules presented by Rechtschaffen and Kales (the R&K rules) [1], human sleep is comprised of wake, rapid-eye-movement (REM) sleep and four non-REM (NREM) sleep stages S1–S4. S1 and S2 are usually grouped as "light sleep" and S3 and S4 correspond to slow-wave sleep (SWS) or "deep sleep" [2]. The gold standard for nocturnal sleep assessment is overnight polysomnography (PSG) which is typically collected in a sleep laboratory. With PSG, sleep stage is manually scored on each 30-s epoch throughout the night by trained sleep experts, forming a sleep hypnogram [1]. PSG recordings usually contain multiple bio-signals such as electroencephalography (EEG), electrocardiography (ECG), electrooculography (EOG), electromyography (EMG), respiratory effort, and blood oxygen saturation.

Respiratory information has been widely used for objectively assessing human nocturnal sleep [3–5]. Detecting sleep stages overnight is beneficial to the interpretation of sleep architecture or monitoring of sleep-related disorders [6,7]. Cardiorespiratory-based automated sleep stage classification has been increasingly studied in recent years [8–12]. Some of those studies only made use of respiratory activity because, when comparing with it cardiac activity is relatively more difficult to be captured reliably in an unobtrusive manner [10,11]. For respiratory activity, in comparison with the breathing ventilation acquired with traditional devices such as nasal prongs or face mask [13], respiratory effort can be obtained in an easier and more noninvasive or unobtrusive way, e.g., using a respiratory inductance plethysmography (RIP) sensor [14], an infrared (IR) camera [15], or a pressure sensitive bed-sheet [16].

Several parameters have been derived from respiratory effort signals for sleep analysis including respiratory frequency, powers of different respiratory spectral bands [8], respiratory self-similarity

* Corresponding author at: Signal Processing Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, Den Dolech 2, 5612 AE Eindhoven, The Netherlands. Tel.: +31 6 8647 1335.
E-mail addresses: xi.long@philips.com, x.long@tue.nl, xi.long.ee@gmail.com (X. Long).

http://dx.doi.org/10.1016/j.bspc.2014.08.001
1746-8094/

[11], regularity [17] etc. These parameters are usually called "features" in the tasks of epoch-by-epoch sleep stage classification. In addition, it has been reported that the respiratory amplitude (e.g., depth and volume) differs between sleep stages [4]. For instance, the "respiratory depth" is more regular and the tidal volume, minute ventilation, and inspiratory flow rate are significantly lower during REM sleep than during NREM sleep (particularly during deep sleep) [18,19]. To the authors' knowledge, these characteristics that express different physiological properties across sleep stages have not been explicitly elaborated and quantified for applications of sleep stage classification. We therefore exploit these characteristics by analyzing respiratory effort signal envelope and area. Features quantifying these characteristics are motivated to be designed which are expected to in turn help separate different sleep stages.

It is assumed that the information about respiratory depth or volume is obtainable from the respiratory effort signal. For instance, the signal (upper and lower) envelopes and area should correspond to respiratory depth and volume, respectively. In fact, respiratory effort has often been used as a surrogate of tidal volume since it is obtained by measuring motions of rib cage or abdominal with, e.g., RIP [14]. However, Whyte et al. [20] argued that this assumption does not always hold, particularly when a sleeper changes his/her posture along with body movements during sleep. This is because the respiratory effort amplitude might be affected by body movements as the sensor position may shift and/or the sensor may be stretched. This will cause an uneven comparison of the signal amplitude before and after body movements, yielding errors when computing the feature values. In order to provide a more accurate estimate of respiratory depth and volume from respiratory effort signal, we must calibrate the signal by means of body movements. They can be quantified by analyzing the artifacts of respiratory effort signal (often inline with body movements) using a dynamic time warping (DTW)-based method [11]. DTW is a signal-matching algorithm that quantifies an optimal non-linear alignment between two time series allowing scaling and offset [21]. Our previous work [11] has proposed a DTW measure to effectively capture body motion artifacts by measuring self-similarity of respiratory effort. This measure has been successfully used as a feature for classifying sleep and wake states in that work. Therefore, we simply adopted this measure to detect motion artifacts modulated by body movements in respiratory effort signals. Using the DTW-based method enables the exclusion of an additional sensor modality (e.g., actigraphy) specifically used for detecting body movements.

The address of this paper is exclusively on investigating a set of novel features that can characterize respiratory amplitude in different aspects with the ultimate goal of improving sleep stage classification performance. Previous studies have shown that linear discriminant (LD) is an appropriate algorithm in sleep stage classification [6,8,22]. Likewise, we simply adopted an LD classifier. Preliminary results of this work in classifying REM and NREM sleep have been previously published [23].

2. Materials and methods

2.1. Subjects and data

Data of 48 healthy subjects (21 males and 27 females) in the SIESTA project (supported by European Commission) [24] were included in our data set. The subjects had a Pittsburgh Sleep Quality Index (PSQI) of no more than 5 and met several criteria (no shift work, no depressive symptoms, usual bedtime before midnight, etc.). All the subjects signed an informed consent form prior to the study, documented their sleep habits over 14 nights, and underwent overnight PSG study for two consecutive nights (on day 7 and day 8) in sleep laboratories. The PSG recordings collected on day 7 were used for analyses, from which the respiratory effort signals (sampling rate of 10 Hz) were recorded with thoracic inductance plethysmography.

Table 1
Summary of subject demographics and some sleep statistics ($N = 48$).

Parameter	Mean ± SD	Range
Sex	21 males and 27 females	
Age (years)	41.3 ± 16.1	20–83
BMI[a] ($kg\,m^{-2}$)	23.6 ± 2.9	19.1–31.3
TRT[b] (h)	7.8 ± 0.4	6.6–8.6
Wake, W (%)	12.9 ± 6.1	1.2–24.5
REM sleep, R (%)	19.0 ± 3.3	15.3–26.5
NREM sleep, N (%)	68.1 ± 4.9	56.1–76.3
Light sleep, L (%)	53.6 ± 5.5	42.7–66.7
Deep sleep, D (%)	14.5 ± 4.8	5.3–28.5

[a] Body mass index.
[b] Total recording time.

Sleep stages were manually scored on 30-s epochs as wake, REM sleep, or one of the NREM sleep stages by sleep clinicians based on the R&K rules. For sleep stage classification epochs were labeled as four classes W (wake), R (REM sleep), L (light sleep), and D (deep sleep), or three classes W, R, and N (NREM sleep).

From the data used in this study the subject demographics and some sleep statistics [mean ± standard deviation (SD) and range] are summarized in Table 1.

2.2. Signal preprocessing

The raw respiratory effort signals of all subjects were preprocessed before feature extraction. They were filtered with a 10th order Butterworth low-pass filter with a cut-off frequency of 0.6 Hz for the purpose of eliminating high frequency noise. Afterwards the baseline was removed by subtracting the median peak-to-trough amplitude. To locate the peaks and troughs, we identified the turning points simply based on sign change of signal slope and then corrected the falsely detected 'dubious' peaks and troughs (1) with too short intervals between peak and trough pairs where the sum of two successive intervals is less than the median of all intervals over the entire recording and (2) with two small amplitudes where the peak-to-trough difference is smaller than 15% of the median of the entire respiratory effort signal. These methods were validated by comparing automatically detected results with manually annotated peaks and troughs and an accuracy of ~98% was achieved.

2.3. Existing respiratory features

A pool of 14 existing features extracted from the respiratory effort signal has been used in previous studies for sleep stage classification. In the time domain, the mean and SD of breath lengths (L_m and L_{sd}) and the mean and SD of breath-by-breath correlations (C_m and C_{sd}) were calculated [6]. In the frequency domain, we extracted features based on the respiratory effort spectrum for each epoch where the spectrum was estimated using a short time Fourier transform (STFT) with a Hanning window. From the spectrum the dominant frequency (F_r) in the range of 0.05–0.5 Hz (estimated as the respiratory frequency) and the logarithm of its power (F_p) were obtained [6]. We also took the logarithm of the spectral power in the very low frequency band between 0.01 and 0.05 Hz (VLF), low frequency band between 0.05 and 0.15 Hz (LF), and high frequency band from 0.15 to 0.5 Hz (HF) and the ratio between LF and HF spectral powers (LF/HF) [6,8]. Furthermore the standard deviation of respiratory frequency over 5 epochs (F_{sd}) was computed [8]. Non-linear features consist of self-similarity measured

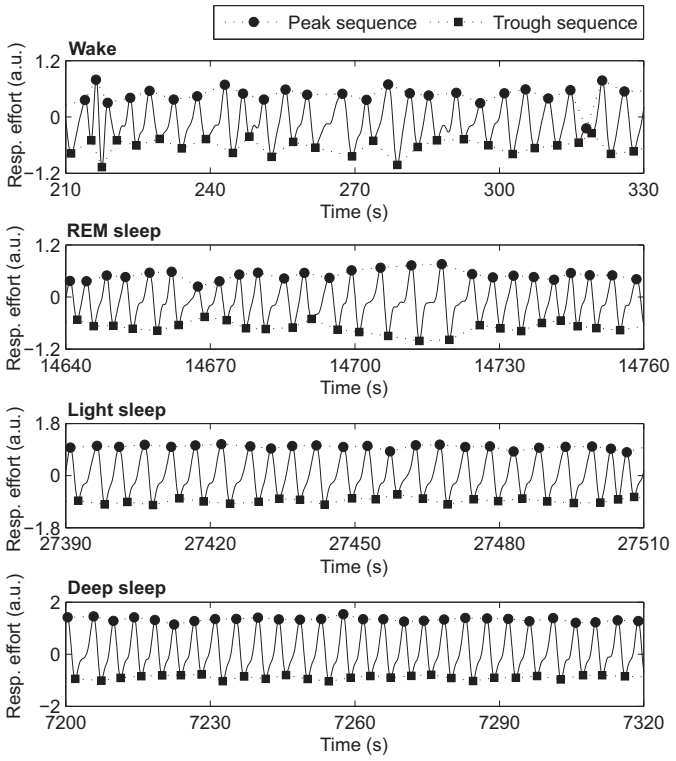

Fig. 1. A typical example of a 2-min (or 4-epoch) respiratory effort signal in wake, REM sleep, light sleep and deep sleep. The peaks and troughs are represented by filled circles and squares, respectively.

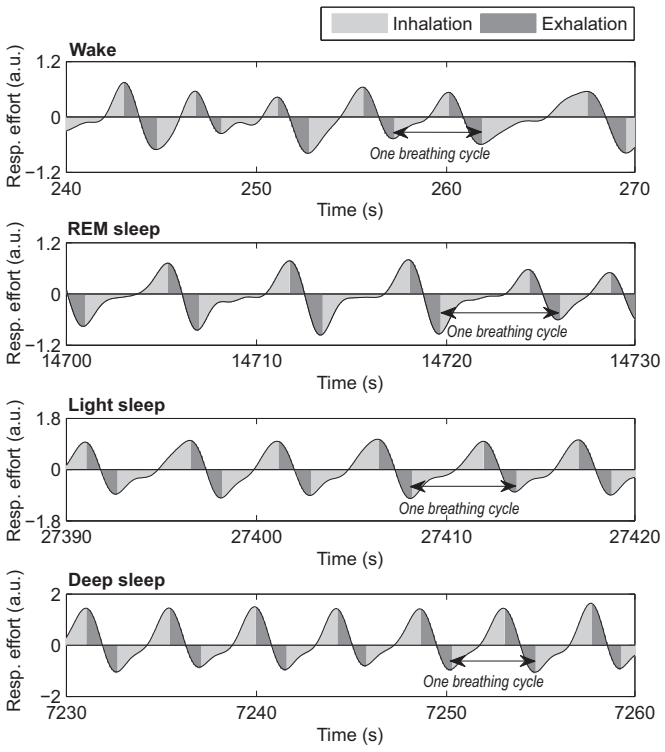

Fig. 2. A typical example of a 30-s (or one-epoch) respiratory effort signal in wake, REM sleep, light sleep and deep sleep. The areas between the curves and the baseline are filled in light gray (inhalation) and dark gray (exhalation). Examples of one breathing cycle period in different sleep stages are indicated.

between each epoch of interest and the other epochs by means of dynamic time and frequency warping (S_{dtw} and S_{dfw}) [11] and signal regularity estimated by sample entropy (R_{se}) [17]. The latter was implemented with the PhysioNet toolkit *sampen* [25].

2.4. Respiratory amplitude features

2.4.1. Analysis of respiratory effort amplitude

Fig. 1 illustrates four short segments of a respiratory effort signal during different sleep stages. It is observed that the envelopes formed by the peak and trough sequences of the signal during wake and REM sleep, when compared with that during light and deep sleep: (1) are more 'irregular'; (2) have generally lower absolute mean or median; and (3) have larger variance. In addition, as illustrated in Fig. 2, we also considered the respiratory effort 'area' comprised between the respiratory effort amplitude and its mean value (zero in the example). As explained, this area should correlate with respiratory volume to a certain extent, which differs across sleep stages. Relying on these observations, several new respiratory amplitude features were explored in two aspects, namely respiratory depth-based and volume-based features.

2.4.2. Depth-based features

A total of five depth-based features were extracted from the peak and trough sequences (i.e., upper and lower envelopes) of the respiratory effort signal. The amplitudes of these peaks and troughs should include the information in regard to respiratory depth. Let us consider $p = p_1, p_2, \ldots, p_n$ and $t = t_1, t_2, \ldots, t_n$ the peak and trough sequences from a window of 25 epochs or 12.5 min centered at the epoch under consideration, containing n peaks and troughs, respectively. We thus computed the standardized median of the peaks (and troughs) by dividing the median by their interquartile range (IQR, the difference between the 3rd and the 1st quartile), such that

$$P_{sdm} = \frac{\text{median}(p_1, p_2, \ldots, p_n)}{\text{IQR}(p_1, p_2, \ldots, p_n)}, \quad (1)$$

$$T_{sdm} = \frac{\text{median}(t_1, t_2, \ldots, t_n)}{\text{IQR}(t_1, t_2, \ldots, t_n)}. \quad (2)$$

These two features consider the mean respiratory depth and its variability at the same time in terms of inhalation (for peaks) and exhalation (for troughs). Note that the period length of 25 epochs was chosen to maximize the average discriminative power (see Section 2.7.2) of all respiratory amplitude features in separating wake, REM sleep, light sleep, and deep sleep.

In order to examine how regular the envelopes are, we used the non-linear sample entropy measure, which has been broadly used in quantifying regularity of biomedical time series [17]. Now considering a time series with n data points $u = u_1, u_2, \ldots, u_n$, let $v(i) = u_i, u_{i+1}, \ldots, u_{i+m-1}$ $(1 \le i \le n-m+1)$ be a subsequence of u, where the window length m is an positive integer and $m < n$. Then for each i, we have $B_{i,m}(r) = (n-m+1)^{-1}\eta(r)$, in which $\eta(r)$ is the number of j such that $d_m[v(i), v(j)] \le r$ $(1 \le j \le n-m, j \ne i)$ where the distance metric d_m between two subsequences $v(i)$ and $v(j)$ is given by $d_m[v(i), v(j)] = \max |u_{i+l} - u_{j+l}|$ for all $l = 0, 1, \ldots, m-1$. For a higher dimension $m+1$, we have $A_{i,m}(r)$. Then the sample entropy of the time series u is defined by

$$SE = -\ln\left[\frac{A^m(r)}{B^m(r)}\right], \quad (3)$$

where

$$A^m(r) = \frac{1}{n-m}\sum_{i=1}^{n-m} A_{i,m}(r), \quad (4)$$

$$B^m(r) = \frac{1}{n-m}\sum_{i=1}^{n-m} B_{i,m}(r). \quad (5)$$

Similarly, the sample entropy measures of the peak and trough sequences are

$$P_{se} = -\ln\left[\frac{A^m_{peak}(r)}{B^m_{peak}(r)}\right], \quad (6)$$

$$T_{se} = -\ln\left[\frac{A^m_{trough}(r)}{B^m_{trough}(r)}\right], \quad (7)$$

in which r is the tolerance that usually takes the value of 0.1–0.25 SD of the peak or the trough sequence and m takes a value of 1 or 2 for the sequence of length n larger than 100 data points [17,26]. In our study, r of 0.20 SD of the sequence and m of 2 were experimentally chosen to maximize the discriminative power of the two features.

Additionally, the median of peak-to-trough differences expresses the range of inhale and exhale depths. It was computed as

$$PT_{diff} = \text{median}[(p_1 - t_1), (p_2 - t_2), \ldots, (p_n - t_n)]. \quad (8)$$

2.4.3. Volume-based features

A total of seven volume-based features were extracted from the respiratory effort signal. They should reflect certain properties of respiratory volume. The respiratory effort signal (sampled at 10 Hz) over a window of 25 epochs or 12.5 min centered at the epoch of interest is expressed as $s = \{s_1, s_2, \ldots, s_x, \ldots, s_M\}$ ($x = 1, 2, \ldots, M$), where M is the number of sample points in this period. Suppose that Ω^{br}_k is the kth breathing cycle in the epoch where there are in total K consecutive breathing cycles ($k = 1, 2, \ldots, K$). Then the corresponding kth inhalation and exhalation periods are Ω^{in}_k and Ω^{ex}_k, respectively. As illustrated in Fig. 2, a breathing cycle is the period between two consecutive troughs and thereby the inhalation and exhalation periods in this breathing cycle are separated by the peak in between these two troughs. We first computed the median respiratory volume (expressed by respiratory effort area) measured during breathing cycles (V_{br}), inhalation periods (V_{in}), and exhalation periods(V_{ex}) for each epoch, such that

$$V_{br} = \text{median}\left(\sum_{s_x\in\Omega^{br}_1} s_x, \sum_{s_x\in\Omega^{br}_2} s_x, \ldots, \sum_{s_x\in\Omega^{br}_K} s_x\right), \quad (9)$$

$$V_{in} = \text{median}\left(\sum_{s_x\in\Omega^{in}_1} s_x, \sum_{s_x\in\Omega^{in}_2} s_x, \ldots, \sum_{s_x\in\Omega^{in}_K} s_x\right), \quad (10)$$

$$V_{ex} = \text{median}\left(\sum_{s_x\in\Omega^{ex}_1} s_x, \sum_{s_x\in\Omega^{ex}_2} s_x, \ldots, \sum_{s_x\in\Omega^{ex}_K} s_x\right). \quad (11)$$

In addition, we computed the median respiratory "flow rate" (expressed by the respiratory effort area over time) during breathing cycles (FR_{br}), inhalation periods (FR_{in}), and exhalation periods (FR_{ex}), such that

$$FR_{br} = \text{median}\left(\frac{1}{\tau^{br}_1}\sum_{s_x\in\Omega^{br}_1} s_x, \frac{1}{\tau^{br}_2}\sum_{s_x\in\Omega^{br}_2} s_x, \ldots, \frac{1}{\tau^{br}_K}\sum_{s_x\in\Omega^{br}_K} s_x\right), \quad (12)$$

$$FR_{in} = \text{median}\left(\frac{1}{\tau^{in}_1}\sum_{s_x\in\Omega^{in}_1} s_x, \frac{1}{\tau^{in}_2}\sum_{s_x\in\Omega^{in}_2} s_x, \ldots, \frac{1}{\tau^{in}_K}\sum_{s_x\in\Omega^{in}_K} s_x\right), \quad (13)$$

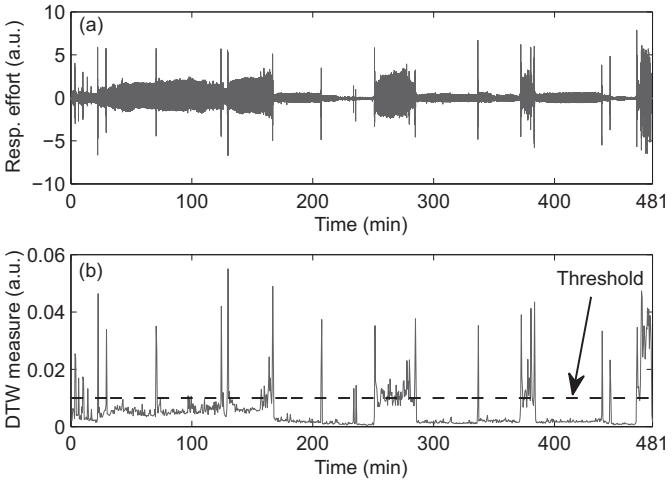

Fig. 3. An example of (a) an overnight respiratory effort signal and (b) the corresponding epoch-based DTW measure, where the threshold (0.01) for identifying epochs with body movements is indicated.

$$FR_{ex} = \text{median}\left(\frac{1}{\tau^{ex}_1}\sum_{s_x\in\Omega^{ex}_1} s_x, \frac{1}{\tau^{ex}_2}\sum_{s_x\in\Omega^{ex}_2} s_x, \ldots, \frac{1}{\tau^{ex}_K}\sum_{s_x\in\Omega^{ex}_K} s_x\right), \quad (14)$$

in which τ^{in}_k and τ^{ex}_k are the kth inhalation and exhalation time (unit: 100 ms)

$$\tau^{in}_k = \max_{s_x\in\Omega^{in}_k}(x) - \min_{s_x\in\Omega^{in}_k}(x), \quad (15)$$

$$\tau^{ex}_k = \max_{s_x\in\Omega^{ex}_k}(x) - \min_{s_x\in\Omega^{ex}_k}(x), \quad (16)$$

and accordingly the time of the kth breathing cycle is given by

$$\tau^{br}_k = \tau^{in}_k + \tau^{ex}_k. \quad (17)$$

The ratio of the inhalation and the exhalation flow rate FR_{in} and FR_{ex} was finally computed as

$$RT_{fr} = \frac{FR_{in}}{FR_{ex}}. \quad (18)$$

2.4.4. Signal calibration by body movements

As mentioned, the respiratory amplitude features are sensitive to body motion artifacts. We thus should calibrate the respiratory effort signal before computing these features. This was done by calibrating each signal segment to have zero mean and unit variance between any two epochs detected as with body movements. As mentioned in Section 1, a DTW-based method measuring the respiratory similarity between each epoch and its adjacent epochs using DTW distance [21] was applied to estimate the body movements. For the details of computing the DTW measure we refer to our previous work [11]. Here the epochs were identified as with body movements if their DTW measures (expressing body motion artifacts) are larger than a threshold. A threshold of 0.01 was experimentally found to be adequate for this purpose. Fig. 3 compares an overnight preprocessed respiratory effort signal with the corresponding epoch-based DTW measure from a subject where the peaks (reflecting body movements) are well aligned in time axis.

2.5. Subject-specific feature normalization

Following the feature extraction procedure as described above, we performed a subject-specific Z-score normalization for each feature. It was done per subject/recording by subtracting the mean of feature values and dividing by their standard deviation. This

allows for reducing physiological and equipment-related variations from subject to subject, thereafter enhancing the discrimination between sleep stages.

2.6. Classifier

An LD classifier was used for sleep stage classification in this study. With LD, the prior probabilities of different classes (i.e., sleep stages) have been observed to change over time. To exploit this change, we calculated a time-varying prior probability for each epoch by counting the relative frequency that specific epoch index was labeled as each class [6,8,22].

2.7. Experiments and evaluation

2.7.1. Cross validation

A 10-fold cross validation (10-fold CV) was conducted in our experiments. The subjects were first randomly divided into 10 subsets, yielding 8 subsets with 5 subjects each and 2 subsets with 4 subjects each. During each iteration of the 10-fold CV procedure, data from 9 subsets were used to train the classifier and the remaining one was used for testing. After CV, classification results obtained for each subject in each iteration's testing set were collected and performance metrics (averaged or pooled over all subjects) were computed to evaluate the classifier.

2.7.2. Feature evaluation and ranking

We first compared the values of the new respiratory amplitude features in different sleep stages to see whether they are statistically different between sleep stages. This serves to understand their feasibility to detect sleep stage at first glance. For each of them, an unpaired Mann–Whitney test (two-sided) was applied to examine the significance of difference.

To assess the discriminative power or class separability of each single feature in separating different classes, the information gain (IG) [27] metric was employed. IG describes the change in information entropy caused by knowing the informative feature values. A higher discriminative power of a feature is reflected by a larger IG value, vice versa. In this study the discriminative power of the new features (in separating wake, REM sleep, light sleep, and deep sleep) with and without calibrating the respiratory effort signal and with and without performing subject-specific normalization were compared. To examine which sleep stage they are able to detect best, we compared their IG values (after signal calibration and feature normalization) in discriminating between each stage and all the other stages as a whole. The new features in combination with the existing features were ranked by IG which serves to select features.

During each 10-fold CV iteration, features were first ranked by means of the discriminative power (measured by IG) in a descending order based on the associated training set. Afterwards, a certain number of top-ranked features were selected. With this approach, we would get 10 feature subsets for all the 10 iterations. To compare the classification performance using different number of features, we plot the performance metric versus the number of selected features and then report the best results. Note that the feature ranking and thus the selected features may change during each iteration of the cross validation. We allowed for this during our experiments since we found that the feature rankings in different iterations were similar for the relatively large-sized training data sets (with 43 or 44 overnight recordings) used in this study.

2.7.3. Classification performance evaluation

We evaluated the performance of several sleep stage classification tasks. They are (1) two multiple-stage classification tasks: WRLD (classification of W, R, L, and D) and WRN (classification of W, R, and N); and (2) four detection tasks: W, R, D, and N (binary classification between each of them versus all the other stages).

To evaluate the performance of classifiers, conventional metric of overall accuracy was considered. However, the high class imbalance makes this metric less appropriate. For instance, the wake epochs account for an average of only 12.9% of all the epochs throughout the night while the light sleep constitutes 53.6% of the night. The Cohen's Kappa coefficient of agreement [28] which has often been used in the area of sleep stage classification is considered to be a better criterion for this problem. By factoring out chance agreement, it is not sensitive to class imbalance. By these means, it offers a better understanding of the general performance of the classifier in correctly identifying different classes. For the binary classification tasks, we chose the classifier decision-making threshold leading to the maximum pooled Kappa and therefore with this threshold the mean and SD of the overall accuracy and Kappa over all subjects were computed.

For each classification task, the 10-fold CV using the LD classifier was conducted with the feature sets comprising the existing pool of 14 respiratory features (set "exist") and the combination of the existing features and the new respiratory amplitude features (set "all"). In addition, we also compared the classification results obtained using features with and without performing subject-specific (Z-score) normalization. A paired Wilcoxon signed-rank test (two-sided) was applied to test the significance of difference between classification performances.

3. Results

As shown in Fig. 4, the respiratory amplitude features were found to significantly differ across sleep stages. This means that the information regarding respiratory depth and volume estimated from respiratory effort, which are indicators of some properties of respiratory physiology, is not independent of sleep stages and thus it can be in turn used to classify sleep stages.

Fig. 5 compares their discriminative power in separating wake, REM sleep, light sleep and deep sleep with and without respiratory signal calibration (by means of body motion artifacts) and subject-specific feature normalization. Mostly, by calibrating the respiratory effort and normalizing the features per subject, the IG values of these new features were increased. The discriminative powers of all the 26 respiratory features for different classification tasks are presented in Fig. 6. We note that the respiratory amplitude features rank higher than most existing features for multiple-stage classifications and NREM sleep detection. P_{sdm} and T_{sdm} (reflecting the variability of depth) perform better in detecting deep sleep; P_{se} and T_{se} (reflecting the regularity of respiratory depth) have a relatively larger power in distinguishing between wake and sleep. It can be seen that the volume-based features (with an exception of RT_{fr}) have higher discriminative powers in detecting REM sleep.

Fig. 7 illustrates the average Cohen's Kappa coefficient versus the number of features (ranked and selected by IG values) used for different classification tasks. For most tasks the classification performance obtained using the feature set "all" is always better than that obtained using the feature set "exist" when the number of selected features is larger than a certain value. The overall accuracy and Kappa coefficient with the number of selected features yielding maximum Kappa are summarized in Table 2. We see that, on the one hand, normalizing the features per subject largely increased the sleep stage classification performance for all the classification tasks. It also shows that, to a certain extent, this method is able to reduce between-subject variability in respiratory physiology (by comparing their SD). On the other hand, combining the existing and the new respiratory amplitude features resulted in significantly improved results except for wake detection. In

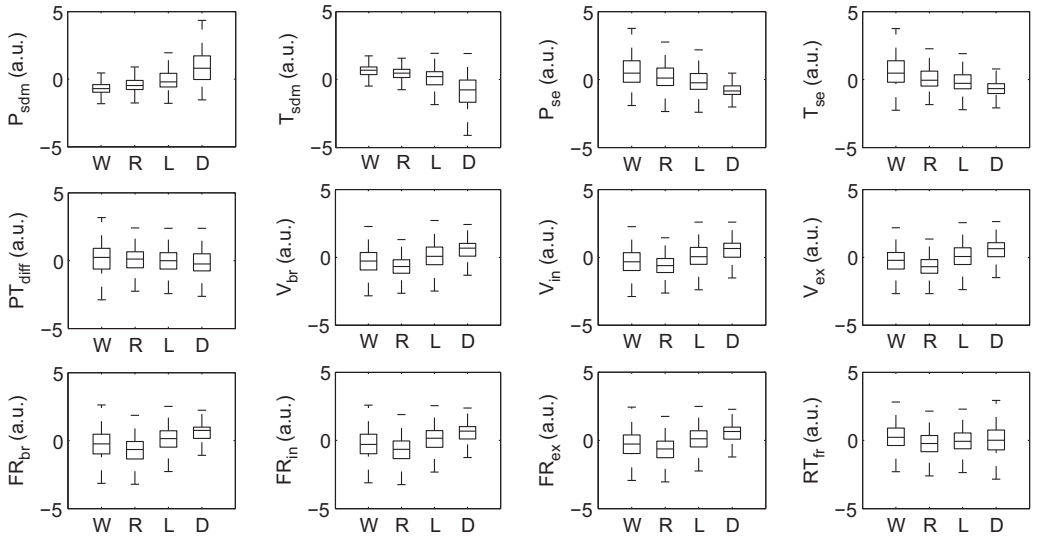

Fig. 4. Boxplots of values of the 12 respiratory amplitude features (with signal calibration and subject-specific normalization) in different classes (W, R, L and D). Outliers are not shown in order to visualize the boxes clearer. The significance of difference was found between each two classes for each feature using an unpaired Mann–Whitney test at $p < 0.01$.

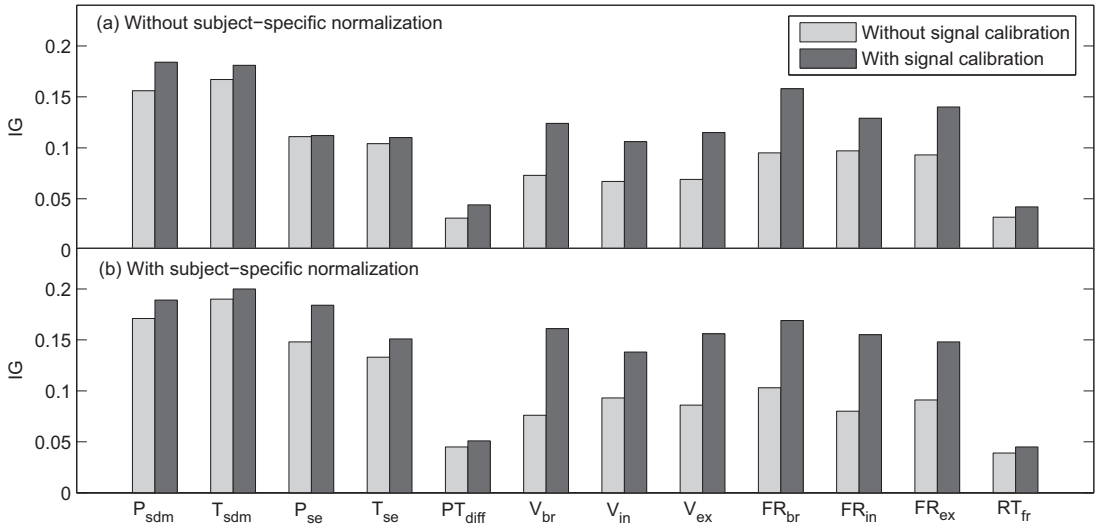

Fig. 5. Comparison of discriminative power (as measured by IG) of all the 12 respiratory amplitude features without and with calibrating the respiratory effort signals for WRLD classification, where the values (a) without and (b) without subject-specific feature normalization are both presented. IG was computed by pooling epochs over all subjects.

Table 2
Summary of sleep stage classification performance (10-fold CV) using feature set "exist" and "all" with and without performing subject-specific feature normalization.

Task	Feat. set	Without subject-specific normalization			With subject-specific normalization		
		# Feat.	Acc. (%)	Kappa	# Feat.	Acc. (%)	Kappa
WRLD classification	Exist	14	58.4 ± 6.8	0.26 ± 0.12	13	61.7 ± 6.9	0.32 ± 0.11
	All	25	59.2 ± 8.6*	0.29 ± 0.14**	24	63.8 ± 8.1***	0.38 ± 0.14***
WRN classification	Exist	14	71.7 ± 7.4	0.32 ± 0.14	13	75.0 ± 6.7	0.41 ± 0.13
	All	25	72.3 ± 8.1*	0.34 ± 0.15*	23	76.2 ± 7.9**	0.45 ± 0.15***
W detection	Exist	6	89.8 ± 6.3	0.49 ± 0.16	10	90.1 ± 4.2	0.50 ± 0.14
	All	9	89.8 ± 6.2^{NS}	0.49 ± 0.16^{NS}	15	90.3 ± 4.1^{NS}	0.51 ± 0.15^{NS}
R detection	Exist	14	79.4 ± 7.8	0.29 ± 0.19	14	82.0 ± 5.6	0.39 ± 0.20
	All	26	79.9 ± 7.6^{NS}	0.31 ± 0.19*	26	82.7 ± 5.8^{NS}	0.44 ± 0.20**
D detection	Exist	12	84.6 ± 4.9	0.26 ± 0.19	10	84.9 ± 4.3	0.33 ± 0.17
	All	8	86.1 ± 4.1**	0.34 ± 0.22**	5	86.1 ± 4.1*	0.43 ± 0.19***
N detection	Exist	13	72.8 ± 10.8	0.40 ± 0.17	14	75.2 ± 8.0	0.44 ± 0.17
	All	23	73.3 ± 11.6^{NS}	0.42 ± 0.19*	25	76.8 ± 8.7*	0.48 ± 0.18**

Note: For each feature set, the results obtained using the selected features leading to maximum Kappa coefficient are reported (see Fig. 7). Significance of difference between the results obtained using feature set "exist" and "all" was examined with a paired two-sided Wilcoxon signed-rank test ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, NS: not significant). For all metrics, significant difference was found between the results obtained with and without subject-specific feature normalization at $p < 0.01$ except for wake detection.

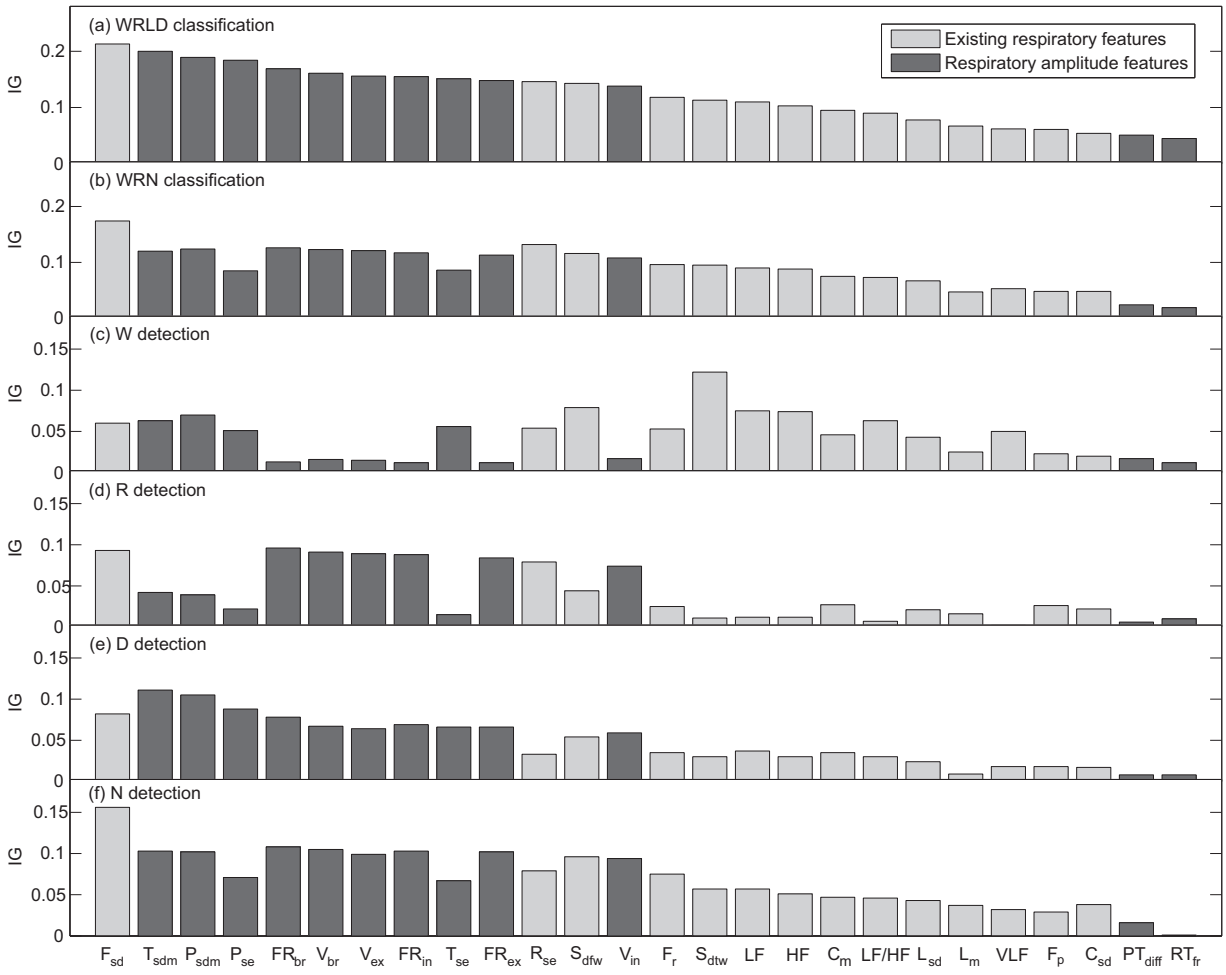

Fig. 6. Discriminative power of all the 26 respiratory features (with signal calibration and subject-specific feature normalization) for (a) WRLD classification, (b) WRN classification, (c) W detection, (d) R detection, (e) D detection, and (f) N detection. The features were ranked by IG (computed by pooling epochs over all subjects) for WRLD classification in a descending order.

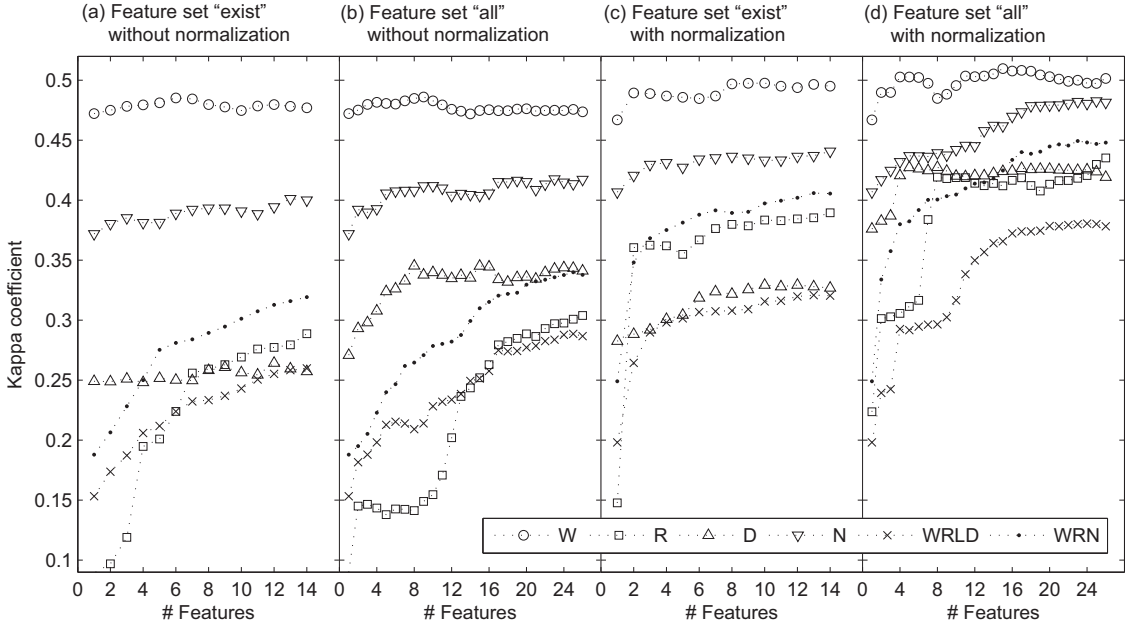

Fig. 7. Kappa coefficient of Sleep stage classification versus the number of selected features ranked by their IG values in a descending order. Results were obtained based on 10-fold CV using feature set (a) "exist" and (b) "all" without subject-specific feature normalization and using feature set (c) "exist" and (d) "all" with subject-specific feature normalization. WRLD: classification of wake, REM sleep, light sleep and deep sleep; WRN: classification of wake, REM sleep and NREM sleep; W: wake detection; R: REM sleep detection; D: deep sleep detection; N: NREM sleep detection.

Table 3
Comparison of multiple-stage classification results with those reported in literature.

Task	First author/year	Signal modalities	# Subjects	# Feat.	Classifier	Acc. (%)	Kappa
WRLD classification	Hedner, 2011 [29]	PAT, PR, OS, AC	227	–	zzzPAT[c]	65.4	0.48
	Isa, 2011 [33]	ECG	16	9	RF	60.3	0.26
	This paper	RE	48	26	LD	63.8	0.38
WRN classification	Redmond, 2007 [8]	ECG, RE	31	30	LD	76.1	0.46
	Mendez, 2010 [30]	BCG[a]	17	46	KNN	72.0	0.42
	Kortelainen, 2010 [31]	BCG[b]	18	4	HMM	79.0	0.44
	Sloboda, 2011 [32]	RE	16	9	NB	~70	–
	Xiao, 2013 [12]	ECG	45	41	RF	72.6	0.46
	This paper	RE	48	26	LD	76.2	0.45

For signal modalities: PAT, peripheral arterial tone; PR, pulse rate; OS, oxyhemoglobin saturation; AC, actigraphy; RE, respiration; BCG, ballistocardiogram measured with bed sensor. For classifier: RF, random forest; LD, linear discriminant; HMM, hidden Markov model; NB, naïve Bayes; KNN, *k*-nearest neighbor.

[a] BCG with cardiac activity and body movement.

[b] BCG with cardiorespiratory activity and body movement.

[c] zzzPAT is a sleep staging algorithm developed by Herscovici et al. [34].

particular, the relatively large improvement in detecting deep sleep epochs (Kappa of 0.43 ± 0.19 versus 0.33 ± 0.17) indicates that the new features can benefit the deep sleep detection most.

Table 3 compares the performance of our sleep stage classifiers (for multiple stages) with those reported in literature. For instance, Hedner et al. [29] presented a Kappa of 0.48 and an overall accuracy of 65.4% in classifying wake, REM sleep, light sleep and deep sleep, which outperform our results but they used more signal modalities such as peripheral arterial tone, pulse rate, oxyhemoglobin saturation and actigraphy. With respect to WRN classification, although Redmond et al. [8] obtained better results compared with our study, they included more signal modalities including cardiac activity. Besides, our results are slightly better than those reported in some other studies, e.g., Kappa of 0.42 by Mendez et al. [30] and Kappa of 0.44 by Kortelainen et al. [31], where they considered ballistocardiogram (BCG) that contains also cardiac information. Nevertheless, when only using respiratory activity, Sloboda et al. [32] achieved an overall accuracy of ~70% (with 9 respiratory features using a naïve Bayes classifier) which is much lower than that presented in this paper.

4. Discussion

The respiratory effort signals were calibrated using the DTW-based method. The DTW measure has been proven to be in association with body movements [11], where a significant Spearman's rank correlation coefficient ($r = 0.32$, $p < 0.0001$) was reported. Further, we obtained a higher correlation ($r = 0.56$, $p < 0.0001$) between the quantified body movements using the DTW-based method (where the DTW measures lower than 0.01 were set to be zero) and activity counts computed using actigraphy based on the data set used in that study. We also tested the sensitivity of the threshold and found that the discriminative power of the respiratory amplitude features did not dramatically change when the threshold was ranging between ~0.005 and ~0.013. To analyze the adequacy of this method for sleep stage classification, we compared the discriminative power as well as the classification performance of these new features between using actigraphy [23] and using the DTW-based method to calibrate the respiratory effort signals. The results are comparable. This suggests that the DTW measure is an adequate estimate of actigraphy for identifying body movements and is therefore effective in mitigating the effect of body motion artifacts on computing the respiratory amplitude features.

As stated in Sections 2.4.2 and 2.4.3, the respiratory amplitude features were computed with a window of 25 epochs (12.5 min). This served to capture the changes of respiratory depth and volume as well as providing reliable regularity measures of peak/trough sequences using sample entropy with sufficient data points. Additionally, we hypothesized that the respiratory effort area can accurately represent breathing tidal volume or ventilation when extracting the respiratory volume-based features. However, this hypothesis is not always acceptable, in particular for subjects who change their posture during sleep [20]. In those cases these features might be inaccurately computed, thus harming classification performance. This challenge should be further studied.

Although the addition of the respiratory amplitude features resulted in enhanced performance in WRLD and WRN classifications (Table 2), the improvements seem relatively modest in general. One explanation is that these new features are correlated with the existing features as discussed before and the additional information is limited. Upon a closer look, we found that the new features contributed more on deep sleep detection than other detection tasks. As a result, this would yield relatively lower performance improvements for multiple-stage classifications since deep sleep only accounts for an average of 14.5% over the entire night. As shown in Table 2, the new features could not help improve wake detection. Actually, the existing features S_{dtw} and S_{dfw} have been shown to be reliable in detecting wake epochs with body movements in our previous study [11]. In this work, to focus more on the respiratory depth and volume properties without being influenced by body movements, we excluded the 'dubious' peaks and troughs (see Section 2.2) where some of them are possibly body motion artifacts which are often indication of wake epochs. Therefore, the new features here might not be able to help detect 'quiet wake' (wakefulness without body movements). Nevertheless, the effect of body movements on the respiratory depth and volume needs to be further studied.

In addition, we observe that the variation of sleep stage classification results between subjects still remains high (see Table 2). For instance, the average Kappa values of WRLD and WRN classifications over all subjects are 0.38 ± 0.14 and 0.45 ± 0.15, respectively. This is mainly caused by large physiological differences between subjects in the way sleep stages are expressed on respiratory features, which naturally leads to difficulties in enhancing the classification performance for some subjects. Therefore, it is still worth investigating methods to reduce the between-subject variability of the features.

In this work we selected features solely based on their discriminative power measured by IG. This approach did not take the correlation or relevance between features into account so that some of them might likely redundant to some extent. On average, the maximum absolute Spearman's rank correlation coefficient $|r|_{max}$ between each new feature and the existing features is 0.35 ± 0.11 (ranging from 0.07 to 0.46 for different new features, $p < 0.01$). For instance, the highest correlation ($r = 0.46$, $p < 0.0001$) occurs between F_{sd} and T_{sdm}, indicating that the variation of respiratory frequency is highly correlated with respiratory depth and its change. Hence, employing feature selectors that aim at reducing

feature redundancy merits further investigation, especially when more features are incorporated.

As presented in Table 3, our methods achieved acceptable sleep stage classification results when using respiratory information alone. Although the results are lower than some other studies, those studies used more signal modalities such as cardiac activity. We therefore anticipate that the classification performance should be further enhanced when combining respiratory and cardiac activity, which will be further studied. Moreover, we only used the simple LD classifier as long as we exclusively focused on analyzing new features for sleep stage classification. Nevertheless, more advanced classification algorithms merit investigation in future work.

5. Conclusions

Respiratory effort amplitude (depth and volume) was analyzed and quantified during nighttime sleep, which has been found to differ across sleep stages. Based on this, 12 novel features that characterize different aspects of respiratory effort amplitude were extracted for automated sleep stage classification. To eliminate the effect of body movements during sleep, respiratory effort signals were calibrated by using a DTW measure which has been shown to correlate with body motion artifacts. By calibrating the signals and normalizing the features for each subject, the discriminative power of the features can be increased. When using only respiratory effort signals, combining the new features proposed in this paper with the existing respiratory features (known in literature) can help significantly improve the performance in classifying and identifying different sleep stages with an exception of wake state detection.

Acknowledgments

The authors would like to thank two anonymous reviewers and Mustafa Radha from Philips Research for their insightful comments.

References

[1] E.A. Rechtschaffen, A. Kales, A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects, National Institutes of Health, Washington, DC, 1968.
[2] M.H. Silber, S. Ancoli-Israel, M.H. Bonnet, S. Chokroverty, M.M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S.A. Keenan, M.H. Kryger, T. Penzel, M.R. Pressman, C. Iber, The visual scoring of sleep in adults, J. Clin. Sleep Med. 3 (2007) 121–131.
[3] T. Penzel, N. Wessel, M. Riedl, J.W. Kantelhardt, S. Rostig, M. Glos, A. Suhrbier, H. Malberg, I. Fietze, Cardiovascular and respiratory dynamics during normal and pathological sleep, Chaos 17 (2007) 015116.
[4] N.J. Douglas, D.P. White, C.K. Pickett, J.V. Weil, C.W. Zwillich, Respiration during sleep in normal man, Thorax 37 (1982) 840–844.
[5] V.K. Somers, M.E. Dyken, A.L. Mark, F.M. Abboud, Sympathetic-nerve activity during sleep in normal subjects, N. Engl. J. Med. 328 (1993) 303–307.
[6] S.J. Redmond, C. Heneghan, Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea, IEEE Trans. Biomed. Eng. 53 (2006) 485–496.
[7] P.A. Estevez, C.M. Held, C.A. Holzmann, C.A. Perez, J.P. Perez, J. Heiss, M. Garrido, P. Peirano, Polysomnographic pattern recognition for automated classification of sleep-waking states in infants, Med. Biol. Eng. Comput. 40 (2002) 105–113.
[8] S.J. Redmond, P. de Chazal, C. O'Brien, S. Ryan, W.T. McNicholas, C. Heneghan, Sleep staging using cardiorespiratory signals, Somnologie 11 (2007) 245–256.
[9] T. Willemen, D. Van Deun, V. Verhaert, M. Vandekerckhove, V. Exadaktylos, J. Verbraecken, S.V. Huffel, B. Haex, J. Vander Sloten, An evaluation of cardio-respiratory and movement features with respect to sleep stage classification, IEEE J. Biomed. Health Inform. 18 (2014) 661–669.
[10] T. Kirjavainen, D. Cooper, O. Polo, C.E. Sullivan, Respiratory and body movements as indicators of sleep stage and wakefulness in infants and young children, J. Sleep Res. 5 (1996) 186–194.
[11] X. Long, P. Fonseca, J. Foussier, R. Haakma, R. Aarts, Sleep, Wake classification with actigraphy and respiratory effort using dynamic warping, IEEE J. Biomed. Health Inform. 18 (2014) 1272–1284.
[12] M. Xiao, H. Yan, J. Song, Y. Yang, X. Yang, Sleep stages classification based on heart rate variability and random forest, Biomed. Signal Process. Control 8 (2013) 624–633.
[13] M. Folke, L. Cernerud, M. Ekstrom, B. Hok, Critical review of non-invasive respiratory monitoring in medical care, Med. Biol. Eng. Comput. 41 (2003) 377–383.
[14] M.A. Cohn, A.S. Rao, M. Broudy, S. Birch, H. Watson, N. Atkins, B. Davis, F.D. Stott, M.A. Sackner, The respiratory inductive plethysmograph: a new non-invasive monitor of respiration, Bull. Eur. Physiopathol. Respir. 18 (1982) 643–658.
[15] Y.M. Kuo, J.S. Lee, P.C. Chung, A visual context-awareness-based sleeping–respiration measurement system, IEEE Trans. Inform. Technol. Biomed. 14 (2010) 255–265.
[16] L. Samy, M.-C. Huang, J. Liu, W. Xu, M. Sarrafzadeh, Unobtrusive sleep stage identification using a pressure-sensitive bed sheet, IEEE Sens. J. 14 (2014) 2092–2101.
[17] J.S. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol. Heart Circ. Physiol. 278 (2000) H2039–H2049.
[18] N.S. Cherniack, Respiratory dysrhythmias during sleep, N. Engl. J. Med. 305 (1981) 325–330.
[19] R.C. Heinzer, F. Series, Normal physiology of the upper and lower airways, in: M.H. Kryger, T. Roth, W.C. Dement (Eds.), Principles and Practice of Sleep Medicine, Saunders Elsevier, St. Louis, MO, 2011, pp. 581–596.
[20] K.F. Whyte, M. Gugger, G.A. Gould, J. Molloy, P.K. Wraith, N.J. Douglas, Accuracy of respiratory inductive plethysmograph in measuring tidal volume during sleep, J. Appl. Physiol. 71 (1991) (1985) 1866–1871.
[21] D.J. Berndt, J. Clifford, Using dynamic time warping to find patterns in time series, in: AAAI Workshop Knowl. Disc. Databases (KDD), 1994, pp. 359–370.
[22] X. Long, P. Fonseca, R. Haakma, R. Aarts, J. Foussier, Spectral boundary adaptation on heart rate variability for sleep and wake classification, Int. J. Artif. Intell. Tools 23 (2014) 1460002.
[23] X. Long, J. Foussier, P. Fonseca, R. Haakma, R.M. Aarts, Respiration amplitude analysis for REM and NREM sleep classification, in: Conf Proc IEEE Eng Med Biol Soc, 2013, pp. 5017–5020.
[24] G. Klosch, B. Kemp, T. Penzel, A. Schlogl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeitlhofer, B. Saletu, W.M. Herrmann, S.L. Himanen, D. Kunz, M.J. Barbanoj, J. Roschke, A. Varri, G. Dorffner, The SIESTA project polygraphic and clinical database, IEEE Eng. Med. Biol. Mag. 20 (2001) 51–57.
[25] D.K. Lake, M.J.R.H. Cao, Sample entropy estimation using sampen, PhysioNet (2012).
[26] D.E. Lake, J.S. Richman, M.P. Griffin, J.R. Moorman, Sample entropy analysis of neonatal heart rate variability, Am. J. Physiol. Regul. Integr. Comp. Physiol. 283 (2002) R789–R797.
[27] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
[28] J.A. Cohen, Coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46.
[29] J. Hedner, D.P. White, A. Malhotra, S. Herscovici, S.D. Pittman, D. Zou, L. Grote, G. Pillar, Sleep staging based on autonomic signals: a multi-center validation study, J. Clin. Sleep. Med. 7 (2011) 301–306.
[30] M.O. Mendez, M. Migliorini, J.M. Kortelainen, D. Nistico, E. Arce-Santana, S. Cerutti, A.M. Bianchi, Evaluation of the sleep quality based on bed sensor signals: time-variant analysis, Conf. Proc. IEEE Eng. Med. Biol. Soc. 2010 (2010) 3994–3997.
[31] J.M. Kortelainen, M.O. Mendez, A.M. Bianchi, M. Matteucci, S. Cerutti, Sleep staging based on signals acquired through bed sensor, IEEE Trans. Inform. Technol. Biomed. 14 (2010) 776–785.
[32] J. Sloboda, M. Das, A simple sleep stage identification technique for incorporation in inexpensive electronic sleep screening devices, in: Conf Proc IEEE Nat Aero Elect Conf, IEEE, 2011, pp. 21–24.
[33] S.M. Isa, I. Wasito, A.M. Arymurthy, Kernel dimensionality reduction on sleep stage classification using ECG signal, Int. J. Comp. Spec. Issue 8 (2011) 115–123.
[34] S. Herscovici, A. Pe'er, S. Papyan, P. Lavie, Detecting REM sleep from the finger: an automatic REM sleep algorithm based on peripheral arterial tone (PAT) and actigraphy, Physiol. Meas. 28 (2007) 129–140.